

Employing a domain ontology to gain insights into user behaviour*

Patricia Kearney[†], Sarabjot Singh Anand[‡], Mary Shapcott[†]

[†]Faculty of Engineering, School of Computing and Mathematics
University of Ulster, Newtownabbey, Co. Antrim, Northern Ireland

Email: {kearney-p3, cm.shapcott }@ulster.ac.uk

[‡]Department of Computer Science, University of Warwick, Coventry, England

Email: s.s.anand@warwick.ac.uk

Abstract

As the Web becomes the de facto window shopping experience for the on-line customer so web personalization is becoming an integral part of many on-line retailers' customer relationship management (CRM) strategy. However Web personalization becomes increasingly difficult as the sheer size and heterogeneous nature of the information available on the web leads to information overload. In this paper we use an on-line movie retailer as a case study. We investigate how web visitor usage data may be combined with semantic domain knowledge to provide a deeper understanding of user behaviour. Our belief is that if we can explain the reasons for the users observed behaviour, we should be able to improve the quality of recommendations generated in the personalization process. In particular we introduce an "impact" measure, based on information theory. The measure captures the influence of a given concept from the domain ontology on user behaviour. We then combine impact measures for each of the concepts within the ontology to create an ontological profile for a user that can be used to personalize future interactions.

1 Introduction

Web personalization is the task of adapting information or services provided by a web site to the needs of a particular user or a set of users [Mobasher *et al.*, 2000]. Although web personalization has become an integral part of the user experience, the sheer size and heterogeneous nature of the information available on the web limits its effectiveness. On-line retailers are increasingly relying on web personalization techniques such as recommender systems, to understand, build and effectively manage their relationship with their visitors and for many, these techniques have become an integral part of the overall CRM strategy. Recommender systems generate various sorts of recommendations for web visitors such as web pages to visit, articles

to read or products to buy. A popular and successfully applied data mining technique used by recommender systems is web usage mining which analyses web server logs containing visitor data. These logs are an important resource for the e-retailer because they provide a record of the click-stream activity of visitors as they navigate their way through the web site, leaving behind their implicit ratings of the various products and services (referred to, collectively, as items) provided by the retailer [Nichols, 1998].

In recent years there has been a large body of research centred on approaches to generating recommendations from the implicit and explicit visitor ratings of items [Montaner *et al.*, 2003; Burke, 2002]. Most of the approaches presented to date use only limited knowledge about the visitor and items. There are a number of reasons for this. Web visitors wish to protect their privacy and are reluctant to provide any explicit data over the web, especially when the value of doing so is not obvious [Berendt *et al.*, 2005]. As regards the items, most content on the web lacks semantics. Although in most instances, a database may be in use, the pages generated by the database do not contain explicit references to the database schema. In addition, the lack of a domain ontology defining the relationships between concepts within the domain of interest, makes it difficult to use information from the database in the recommendation process.

Recent developments in ontology engineering and the promise of the Semantic Web have sparked an interest in developing recommender systems that can use this "deeper" domain knowledge along with user ratings to develop improved recommendations for visitors [Ghani and Fano, 2002; Middleton *et al.*, 2004; Mobasher *et al.*, 2004].

This paper focuses on incorporating domain knowledge into the web usage mining process with the aim of obtaining a deeper understanding of the reasons behind the observed users' behaviour. Our belief is that if we can gauge the reasons for the users' observed behaviour, we should be able to improve the quality of recommendations generated. The result of the mining process is an ontological profile for

* This research was partly funded by the Department of Employment and Learning Northern Ireland

individual visits which can then be used for generating recommendations. We further hypothesise that visitor behaviour cannot be assumed to be consistent across visits as differing visit contexts can affect behaviour. Hence in the paper we analyse visits rather than visitors with the intention of measuring the effect of context on user behaviour.

The layout of the rest of this paper is as follows. Section 2 provides the motivation for the technique presented in the paper. Section 3 discusses our approach to web usage mining in the presence of a domain ontology and the method used to create an ontological profile for an individual visit by a visitor. In Section 4 we present some preliminary results using item semantics and behavioural data obtained from an online movie retailer. We conclude the paper by summarising related research (Section 5), summarising our findings to date and outlining future work to be carried out to extend this research (Section 6).

2 Motivation

Two of the most commonly used approaches to recommendation generation are *content-based* filtering and *collaborative* filtering.

In *content-based filtering* items are recommended because they are similar to items the user has liked in the past. A content-based recommender learns a profile of the user's interests based on the features present in objects the user has rated. For example, text recommendation engines like the newsgroup filtering system NewsWeeder [Lang, 1995] uses the words of their texts as features.

The content-based approach focuses on a single user profile. In contrast, *collaborative filtering* recommends items based on the behaviour of other users who historically have had similar tastes. This technique analyses the visit behaviour of previous web visitors (usually from a ratings database) and compares this with the behaviour of the current visitor. Where visitors with similar tastes are found, recommendations for the current visitor are generated, based on the known preferences of these visitors. This can help overcome the over-specialisation problem of content-based filtering although this approach still suffers significantly from scalability and sparsity problems.

Neither traditional content-based filtering nor collaborative filtering approaches consider domain knowledge to provide a deeper understanding of user behaviour. In content-based filtering, a domain-specific set of features is used to represent the items, although no consideration is given to the relationships between these features. The properties or semantics of the underlying domain are not considered in traditional collaborative filtering.

As an illustration of how a collaborative filtering based recommender system works, consider Table 1 consisting of ratings for five movies by three individuals. Let us assume that we need to recommend movies to the visitor John. A question mark is used to signify that John has not rated the movies "Chicago" and "Sleepless in Seattle". A collaborative filtering system would calculate that Jane has similar

taste to John and hence would conclude that a good recommendation for John would be "Sleepless in Seattle".

Table 1: Movie ratings

| | <i>Bridget Jones Diary</i> | <i>Chicago</i> | <i>Cold Mountain</i> | <i>Sleepless in Seattle</i> | <i>Bridget Jones: Edge of Reason</i> |
|------|----------------------------|----------------|----------------------|-----------------------------|--------------------------------------|
| John | 5 | ? | 4 | ? | 5 |
| Mary | 2 | 2 | 5 | 5 | 2 |
| Jane | 4 | 2 | ? | 5 | 5 |

However, if we were to use a movie ontology such as that used by IMDB [Avanča *et al.*, 2001] we could infer from the two behaviours, that Jane likes movies that belong to the "Romantic Comedy" genre whereas John likes movies with "Renee Zellweger" as the lead actress, no matter what the genre of the movie. Hence, "Chicago", starring Renee Zellweger, rather than "Sleepless in Seattle" would be a more appropriate recommendation for John. It is the impact of the different concepts within the movie ontology on movie selection, by visitors, that needs to be learned and used when making recommendations.

3 Domain Semantics and User Behaviour

In this section we describe our approach to incorporating domain knowledge about items, represented in the form of an ontology, to understand user behaviour. Our ultimate aim is to improve the quality of the recommendations that are generated. Our approach uses the ontology in conjunction with behavioural data to calculate the impact of individual concepts (as defined in the ontology) on the choice of items accessed by the visitor. Once the impact measures have been calculated for a visitor, these are used during recommendation generation to calculate the similarity of the visitor to other visitors.

3.1 Ontology and Instances

An ontology is a conceptualisation of a domain in a human-understandable, but machine-readable format. The ontology consists of entities, attributes, relationships and axioms [Guarino and Giaretta, 1995].

An ontology can be represented as a labelled directed graph, $O = \langle C, E \rangle$, where C is the set of concepts and E is the set of edges between elements of C . We use the notation e_{kj} to represent the edge from concept C_k to C_j .

In an extract from our movie ontology (see Figure 1) we have identified four main classes: movie, genre, actor and director and show these along with some of their attributes. Edges are labelled with the semantic relationships between the concepts (classes and attributes) that the edge is incident on. For instance the edge from concept movie to concept title has the semantic relationship attributeOf and the edge from concept movie to concept actor has an actsIn relationship. Each edge has a cardinality associated with it. For example, the relationship between the movie and actor concepts is one-to-many. This refers to the fact that a

movie may contain many actors. In the same way a director also has a one-to-many relationship. An edge may also have certain attributes of its own. For example, actsIn relationship between the actor and movie concepts may have a weight associated with each actor representing the prevalence of his/her appearance within the film. Although not shown, the Genre concept has its own concept hierarchy with 162 genres identified. Figure 3 shows an extract of this hierarchy.

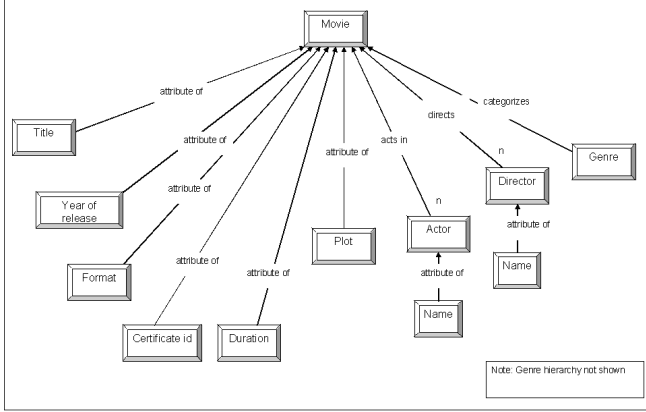


Figure 1: Extract of Movie Ontology

3.2 Mapping URLs onto Concepts

To be able to utilize the domain ontology in conjunction with behavioural data it is essential that a mapping be established from individual URLs on the chosen web site to concepts within the ontology. A web site may contain pages that represent instances of different concepts within the ontology. For example, a movie retailer may provide pages related to individual movies, actors, directors, producers and genre. In our approach we assume that each URL can be identified as being representative of an instance of a particular concept. For database driven web sites, such an assumption is not too restrictive because parameters within the URL often provide the basis for such a mapping to be obtained.

3.3 Ontological Profile

In this section we define an Ontological Profile and describe how it is calculated from behavioural data.

Let $V_p = \langle p_1, p_2, \dots, p_n \rangle$ be a single visit by a visitor, where p_i 's are the individual pages visited by the visitor. Once the pages within a visit had been mapped onto the concepts, the next stage is to generalize from these specific instances of page visits to an Ontological Profile (OP).

Each page, p_i , can be further described using the concepts within the ontology. Let $p_i = \langle x_1, x_2, \dots, x_m \rangle$, where m is the number of concepts in the ontology and x_k is a subset of the instance of C_k ($x_k \subseteq \text{dom}(C_k)$ $1 \leq k \leq m$). Note that some of these x_k 's may be *null*. For example, in the movie domain, if the page is an instance of the actor concept then the values related to the director concept and associated attributes will be *null* as no direct relationship exists between the two concepts. Given this representation of a page, the visit, V_p , can

now be represented in an m -dimensional space, as $\langle d_1, \dots, d_m \rangle$ where, $d_k = \{(v_{ki}, w_{ki}) : v_{ki} \in \text{dom}(C_k) \text{ and } w_{ki} \in [0,1]\}$. We constrain the instance weights w_{ki} such that

$$\sum_{i=1}^{|d_k|} w_{ki} = 1 \quad (1)$$

where $|d_k|$ is the number of unique instances of the concept C_k within the visit V .

We now define the impact, imp_{kj} , for each edge $e_{kj} \in E$, as the impact of concept C_k on C_j inferred from the visitor behaviour within the visit V as $1-h_{kj}$ where h_{kj} is defined as below

$$h_{kj} = \frac{\sum_{i=1}^{|d_k|} w_{ki} \log w_{ki}}{\log |d_k|} \quad (2)$$

The numerator of this equation is the entropy, measuring the disorder or randomness in C_k . The denominator is used to scale the value of the numerator to the interval $[0,1]$. Basing the value of h_{kj} on entropy has the useful feature that when the instances of the concept C_k appear to be randomly occurring within a visit, the value of h_{kj} will be 1. The resulting impact of the concept on what instances of concepts in the adjacency set of C_k appear in the Visit is 0. On the other hand if only one instance of C_k appears in the visit, then the value of h_{kj} will be 0 and the resulting impact value will be 1.

For the purposes of this study we define an Ontological Profile as consisting of the 2-tuple $\langle imp, dist \rangle$. imp is a tuple of weights $\langle imp_{ij} \rangle$ for each $e_{ij} \in E$ and $dist$ is the m -tuple $\langle d_1, \dots, d_m \rangle$ defined above.

Calculating instance weights

The calculation of the instance weights, w_{ki} , depends on the cardinality of the relationship between the concepts, the data type (Nominal or Numeric) of the concept and whether any taxonomy has been defined on the instances of the concept.

To illustrate the instance weight calculation, we use a visit to a movie e-tailer consisting of six pages $\langle p_1, p_2, p_3, p_4, p_5, p_6 \rangle$ as detailed in Figure 2. Pages p_1 and p_2 are instances of two different movie pages. Pages p_3 and p_4 are instances of the same actor page and p_5 and p_6 are instances of the same director page.

Note that the director, Krysstof Kieslowski, is related to four pages in the visit (two movie pages and two director pages). On the other hand while Megan Ward's page is visited twice, neither movie visited were related to her. Intuitively the visit would appear to be strongly influenced by the Director concept while the Actor would appear to have a lesser impact. In this paper aim to quantify this influence or *impact*.

Consider the concepts C_k and C_j such that $e_{kj} \in E$. In the following subsections we describe how the instance weight w_{ki} is calculated for the i th instances of C_k .

Nominal Values with a 1-1 relationship

In this case the value w_{ki} is defined as the frequency with which the value $v_{ki} \in \text{dom}(C_k)$ appears within the visit. If v_{ki} appears n_{ki} times within the n pages visited in visit V then

$$w_{ki} = \frac{n_{ki}}{n} \quad (3)$$

In the example ontology, the Title, Region, Format and Certificate concepts are nominal and have a 1-1 relationship with the concept movie.

$d_{\text{Title}} = \{(The\ Double\ Life\ of\ Veronique, 0.5), (Three\ Colours\ Blue, 0.5)\}$ $\text{Imp}_{\text{Title}} = 0$

$d_{\text{Region}} = \{(0, 0.5), (2, 0.5)\}$ $\text{Imp}_{\text{Region}} = 0$

$d_{\text{Format}} = \{(Widescreen, 0.5), (DVD, 0.5)\}$ $\text{Imp}_{\text{Format}} = 0$

$d_{\text{Certificate}} = \{(15,1)\}$ $\text{Imp}_{\text{Certificate}} = 1$

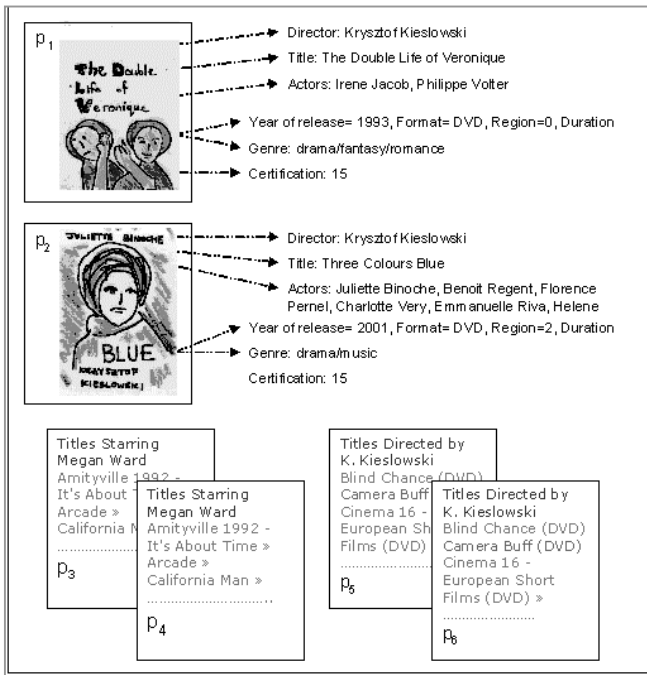


Figure2: Visit example using ontology as described in Figure 1

Nominal Values with a 1-t relationship

Consider a concept C_j that has a one-to-many relationship with concept C_k . This implies that any one instance of concept C_j has t instances of C_k related with it. There may also be some properties of the edge, e_{kj} , that capture further knowledge about the relationship. In the movie context, a movie has a number of actors. This list of actors may have a ranking associated with them that reflects the importance of the role played by the actor in the film. This ranking would be a property of the relationship as opposed to that of the individual concepts. Let $r_{ij} \in [0,1]$ represent the rank of instance i of concept C_k related to instance j of concept C_j . We assume that the rank associated with instances of C_k related to the j th instance of C_j has the following property

$$\sum_{i=1}^t r_{ij} = 1 \quad (4)$$

Once again, this assumption is not severe as a numeric factor can always be scaled to meet this requirement.

Let us assume that the visit V contains u pages that represent instances of the concept C_k and v pages that represent instances of the concept C_j . Let $w \in [0,1]$ be the weight assigned to a visitor accessing a page representing an instance of the concept C_k . We define $r_{ij} = 0$ when the instance i of concept C_k is not related to the j th instance of C_j . Then we calculate w_{ki} as

$$w_{ki} = \frac{wn_{ki} + \sum_{s=1}^v r_{is}}{v + wu}$$

In our example the Actor and Director concepts are nominal and have a 1-t relationship with movie. (5)

$d_{\text{Actor}} = \{(Irene\ Jacob,0.17), (Philippe\ Volter,0.08), (Juliette\ Binoche,0.10), (Benoit\ Regent,0.05), (Florence\ Pernel,0.04), (Charlotte\ Very,0.02), (Emmanuelle\ Riva,0.02), (Helene\ Vincent,0.02), (Megan\ Ward,0.5)\}$ $\text{Imp}_{\text{Actor}} = 0.275$

$d_{\text{Director}} = \{(Krzysztof\ Kieslowski,1)\}$ $\text{Imp}_{\text{Director}} = 1$

Numeric Attribute

For numeric attributes we calculate w_{ki} in the same way as a nominal value with the exception that we scale the w_{ki} to take into account the spread of values within the visit. The scaling constant used for this purpose is

$$s_k = \frac{\max_k^v - \min_k^v}{\max_k - \min_k} \quad (6)$$

where \max_k and \min_k are the min and max values of concept C_k , while \max_k^v and \min_k^v are the min and max values for C_k within the visit V .

The scaling constant used above aims to capture the relative spread of values, within the visit, for the concept. This is a rather simplistic measure and we do not claim that it is necessarily the best measure for this, indeed more expressive statistics exist to measures of spread, the value of which need to be investigated.

In our example, Duration and YearOfRelease are numeric. The scaling constants for these concepts are as follows: $(100-93)/(250-3) = 0.03$ and $(2001-1993)/(2005-1978)=0.29$ respectively.

$d_{\text{Duration}} = \{(93, 0.5), (100,0.5)\}$ $\text{Imp}_{\text{Duration}} = 0.97$

$d_{\text{YearOfRelease}} = \{(2001, 0.5), (1993,0.5)\}$ $\text{Imp}_{\text{YearOfRelease}} = 0.71$

Concept with is_a relationships defined on it

Such a concept will necessarily be nominal, however a distance can be calculated between different instances of the concept. Hence, depending on whether this concept has a 1-to-n or 1-to-1 relationship, we can calculate the w_{ki} value as shown previously for nominal values. However, once again we scale the w_{ki} 's using a measure of the average distance between the instances of C_k within the visit. A number of metrics for measuring distance between two values within a taxonomy have been defined in literature [Ganesan *et al.*,

2003]. We use a measure based on the Lowest Common Ancestor defined as

$$d(i_1, i_2) = 1 - \frac{2 \times \delta(LCA(i_1, i_2))}{\delta(i_1) + \delta(i_2)} \quad (7)$$

where $\delta(v)$ is the depth of the node v in the taxonomy and $LCA(i_1, i_2)$ is the common ancestor of i_1 and i_2 within the is-a hierarchy with maximum depth. In this example the Genre concept has an is_a hierarchy defined on it. A fragment of this hierarchy is shown in Figure 3. Using the method outlined in this section we calculate d_{Genre} and Imp_{Genre} as $d_{Genre} = \{(drama/fantasy/romance, 0.5), (drama/music, 0.5)\}$ $Imp_{Genre} = 0.6$

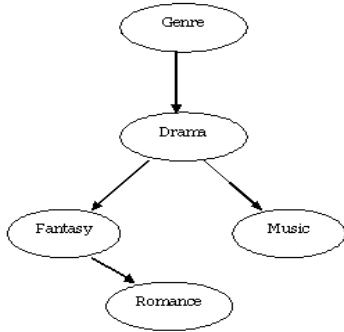


Figure 3: Fragment of Genre Hierarchy

3.4 Measuring Visitor Predictability

We would now like to test our hypotheses that different contexts will result in the same visitor displaying different navigational behaviour. To do this, we cluster the Visit Ontological Profiles generated as described in the Section 3.3.

Let us assume that k clusters are discovered. Each visitor can now be characterized by a vector, V , of size k where $V[i]$ is the number of visits by the visitor that are members of the i th cluster. Given this representation of Visitor behaviour across visits we define the unpredictability of a visitor as

$$p_v = - \sum_{i=1}^k p_i \log p_i \quad (8)$$

where p_i is the probability of a visit by the visitor belonging to the segment i . As P_v is essentially the entropy of visitor behaviour, the larger its value, the more unpredictable is the visitor behaviour.

3.5 Generating a Visitor Ontological Profile

To generate recommendations using these profiles, one of two approaches can be taken. Firstly, the ontological profiles of visits made by a particular visitor can be combined into a Visitor Ontological Profile. Alternatively, anonymous aggregate profiles have been shown to provide useful recommendations without the scalability issues faced with individual user profile based recommendation [Mobasher *et al.*, 2002].

We define a Visitor Ontological Profile using a weighted sum of the Ontological Profiles of the individual visits made

by the visitor u

$$VOP_u = \sum_{i=1}^{|V_u|} \alpha_i \times OP_i \quad (9)$$

where, V_u is the set of visits by visitor u and OP_i is the i th visit by the visitor. The weights, α_i , could be a function of time since the visit so as to degrade the effect of older interactions between the visitor and web site.

3.6 Utilizing Ontological Profiles

Consider a visitor u , with visitor ontological profile $VOP_u = \langle imp_u, dist_u \rangle$, for whom recommendations are to be generated. The distance of visitor u from another visitor v can be calculated as

$$d(u, v) = \sum_{k=1}^m \delta(imp_u^k, dist_u^k, imp_v^k, dist_v^k) \quad (10)$$

where δ measures the distance between two concepts of the ontology. Note that items as well as visitors now have an equivalent representation and hence rather than generating recommendations using a collaborative filtering approach, content based filtering can also be applied, avoiding the new item problem faced by traditional collaborative filtering.

4. Results

4.1 The Data

Two data sets were used to generate the results presented here. One data set consisted of 94,250 instances of the movie ontology stocked by an online movie retailer. For the experiments presented here we used a subset of a movie ontology as shown in Figure 1. The ontology also includes a genre hierarchy e.g. romantic comedy is-a kind of comedy.

To obtain instances of the ontology a spider was developed to extract instances of this ontology from the HTML web pages of the online movie retailer's web site. This information included movie title, actors, genre, plot, directors, duration, year of release etc. We then mapped this knowledge to tables in a database schema, mapping concepts to tables and relationships onto integrity constraints.

The second data set consists of visitor behavioural data collected from web server logs of the same retailer. Currently the data set consists of 193,116 unique visitors, 323,366 unique visits and 1,754,735 movie, actor and director page views. We obtained web usage data from our online movie retailer representing browsing and purchasing behaviour for their customers over a period of time. The data was contained in 122 raw web server log files. These were preprocessed, extracting relevant data to enable identification of unique users and sessions.

Table 2 shows the average impact values for the various concepts within the ontology used in our experiments (see Figure 1). The impact values were generated from 170,000 visits to the online retailer's web site where the visitor visited at least one actor, director or movie page. In general the impact values for actor and director seem quite low. This could possibly be due to the choice of function used by us to

rank actors within a film. In this paper we used the function $f(n) = 1/n$, where n is the order in which the actor appears in the list of actors displayed for the movie on the web site. This is particularly harsh as the actor that is second on the list only gets half the weight given to the first actor in the list. Alternative functions need to be investigated. Region and Format have a low cardinality, each consisting of only two unique instances with approximately 67% of all movie instances belonging to one region and one format, hence the higher impact values were expected.

| Concept | Average Impact | Impact Std. deviation |
|-----------------|----------------|-----------------------|
| Actor | 0.17 | 0.22 |
| Director | 0.27 | 0.4 |
| Genre | 0.45 | 0.45 |
| Title | 0.3 | 0.38 |
| Year Of Release | 0.89 | 0.12 |
| Certificate | 0.58 | 0.35 |
| Format | 0.8 | 0.26 |
| Duration | 0.83 | 0.19 |
| Region | 0.8 | 0.26 |

Table 2: Average Impact of Concepts

The impact value for Duration is not that surprising as the bulk of movies have similar duration values. However there are a number of short films stocked by the retailer and hence the denominator of the scaling constant will be quite large compared to the variation in Duration times within a visit. The Year of release suggests that most visits consist of movie page views of movies released in a fairly narrow time window. On closer examination of the data, the average value of year of release was found to be 1999 with a standard deviation of 2 years. As the behavioural data used in these experiments was that of 2001, we would suggest that the majority of visits consisted of page views associated with “new” movies. Fewer than 1000 visits in the data set have the impact value of Year of Release less than 0.5.

4.2 Mapping URLs onto Concepts

For this study we concentrate on the use of three key concepts. These are the movie, actor and director. The URLs of the web site contained strings that allowed us to easily identify the pages as being instances of one of these three concepts. For example, all movie URLs have a word appearing in the url that differentiate it from pages related to other concepts within the ontology. The URL also contains a unique identifier for the movie associated with the page, that can be used to associate the appropriate semantic data related to the movie with the page.

4.3 Clustering Visit Profiles

Having calculated the Ontological Profiles for each visit, we used the k-Means clustering algorithm to discover any emergent behavioural patterns [Hartigan and Wong, 1979] In the results discussed below, we only considered the impact values for the various concepts. The clustering algorithm discovered some interesting patterns, a selection of these are shown in Table 3. The columns denote the cen-

troids of the segments. We also show the number of visits that fall into each of these segments in the last row of the table. Segments 3, 5 and 6 represent behaviours that are strongly driven by the director concept. Interestingly, the genre is not impacting these visits in Segments 3 and 5; hence the directors that the visitors are showing an interest in are clearly those that do not direct movies in a single genre. Visits that fall into Segment 6 however do display a strong genre impact too.

| Concept | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|-----------------|-------|------|------|------|------|------|------|
| Actor | 0.22 | 0.16 | 0.15 | 0.14 | 0.15 | 0.15 | 0.13 |
| Certificate | 0 | 0.96 | 0 | 0 | 0.42 | 0 | 0.24 |
| Director | 0.78 | 0.2 | 0.99 | 0.0 | 0.99 | 1 | 0.17 |
| Duration | 0.96 | 0.88 | 0.88 | 0.95 | 0.75 | 0.92 | 0.85 |
| Format | 0.01 | 0.99 | 0 | 0 | 0.56 | 0 | 0.46 |
| Genre | 0.73 | 0.06 | 0 | 0.98 | 0.08 | 0.99 | 0.98 |
| Region | 0.02 | 0.99 | 0 | 0 | 0.56 | 0 | 0.47 |
| Title | 0.99 | 0.02 | 0 | 0 | 0.13 | 0 | 0 |
| Year Of Release | 0.92 | 0.97 | 0.9 | 0.91 | 0.86 | 0.9 | 0.87 |
| Visit Count | 22479 | 3295 | 694 | 1115 | 2786 | 689 | 3104 |

Table 3: Patterns discovered in Visit Ontological Profile impact data

With respect to the Genre concept, strong impacts are displayed in Segments 4, 6 and 7. Of these segments we have already discussed Segment 6 and it can be postulated that the high impact of Genre is probably as a result of the Directors of interest to visitors in these visits. However Segments 4 and 7 would appear to be largely driven by Genre alone.

Segment 2 is an interesting segment in that none of the key drivers, namely Genre, Actor and Director seem to be impacting the visits within the segment. We can consider these visitors as browsers or “window shoppers” whose visit context is not to purchase or search for a specific type of item but it is to have a general “look around”. Having said that, the format, region, certificate and year of release have a high impact.

Segment 1, which is by far the largest segment discussed, has a high impact value for the Title concept. Given our current approach to calculating impact for title, it would suggest that these visits are very focused and as such only consist of visitors viewing one movie.

Now let us consider two visitors. The first of these visitors made three visits while the second made seven visits. The behaviour of the first visitor in the first two visits fell into segment 1 while the third visit fell into segment 2. All seven visits by the second visitor belonged to a different segment. Based on the formula in Section 3.4, we can calculate the visitor unpredictability measure for the two visitors as being 0.91 and 2.8, implying that visitor 2 is less predictable than visitor 1. The calculation of visitor unpredictability is not an end in itself though it can be a useful tool when developing individual strategies for personalizing interactions with these visitors in the future. For example,

there is little value in recommending items to a highly unpredictable visitor based on a profile generated from previous visits by the visitor.

4.4 Recommendation Generation

Consider the following two visits. Visit 1, consists of the following two movies {The Jericho Mile, Manhunter} while Visit 2 consists of the two movies {Manhunter, 187}.

Traditional user-based collaborative filtering would recommend 187 during Visit 1 based on the visit's similarity to Visit 2. However, the Ontological Profile of these visits will be very different. In particular, the first visit is impacted by the director (Michael Mann) while the second one is impacted by the movie Genre (Action/Adventure: General). Given this additional knowledge, more appropriate recommendations may be "The Insider" (Genre: Drama:General) or "L.A. Takedown" (Genre: Action/Adventure: General) as these are both Michael Mann films with Genres the same as "The Jericho Mile" and "Manhunter" respectively.

Now consider another visit, Visit 3, consisting of {The Last of the Mohicans, The Insider, L.A.Takedown, Haggard}. This visit is also impacted strongly by the director even though the impact value will not be as high as that of Visit 1 as the movie "Haggard" has not been directed by Michael Mann.

Using the impact values as part of the similarity calculation as described in section 3.6, will result in Visit 2 not being considered similar to Visit 1. However, Visit 3 will be considered as similar to Visit 1 and the movies "The Last of the Mohicans", "The Insider" and "Haggard" may be recommended. Note that "The Last of the Mohicans" belongs to a Genre (Action/Adventure: Romantic) that the visitor in Visit 1 has not shown an interest in, however, the impact of director overrides the difference in Genre, as current behaviour suggests that Genre is not impacting behaviour within Visit 1.

5. Related Work

Recently there has been an interest in using deeper domain knowledge, often represented in the form of an ontology, as part of the recommendation process. The reasons for using the domain knowledge may vary from generating more accurate recommendations to solving issues such as sparsity and scalability, faced by the traditional collaborative filtering approach. In this section we discuss some of this research.

Mobasher *et al.* proposed the use of semantic knowledge about items to enhance item-based collaborative filtering [2004]. Their approach is to represent the semantic knowledge about an item as a feature vector and calculate the similarity based on this information to other items. This item-similarity is then combined with rating similarity to get an overall measure of item similarity which is used to predict the rating by a user of a currently unrated item. This work differs from the approach presented in this paper is a number of ways. Firstly, we aim to measure the underlying impact of various concepts with the domain ontology on the behaviour of a visitor within an individual visit with the

intention of studying variances in behaviour of an individual visitor across multiple visits to a web site. We postulate that the differing behaviour is a result of differing visit context. The intention is to capture not only the semantics and ratings of items but to also incorporate the context of a user visit into the recommendation process. Secondly, we do build user profiles using the semantic knowledge about items and hence our approach is closer to collaborative and content-based hybrid systems rather than the item-based collaborative approach in the paper by Mobasher *et al.*

Dai and Mobasher [2003] provide a framework for integrating domain knowledge with web usage mining for user based collaborative filtering. They highlight that semantics can be integrated at different stages of the knowledge-discovery process. The focus of their research is however on building aggregate profiles rather than ontological profiles for individual visits by a visitor.

Cho and Kim [2004] apply a product taxonomy with web usage mining to reduce the dimensionality of the rating database when searching for nearest neighbours while Niu, Yan *et al.* [2002] build customer profiles based on product hierarchy in order to learn customer preferences. The focus here is more on dimensionality reduction and hence on scalability rather than obtaining an understanding of user behaviour with respect to the underlying domain.

Middleton *et al.* use an ontological profile for a user within their research paper recommendation system, Quick-Step [2004]. The profile is based on a topic hierarchy alone i.e. they only have is_a relationships within the profile. They also attempt to use externally available ontologies based on personnel records and user publications to address the cold-start problem for their recommendations system. The existence of such additional knowledge, while applicable in their specific application domain, cannot however be assumed in a general e-tailer scenario.

Haase *et al.* create semantic user profiles from usage and content information to provide personalized access to bibliographic information on a Peer-to-Peer bibliographic network [2004]. The semantic user profile consists of the expertise, recent queries, recent relevant instances and a set of weights for the similarity function. The user-defined weights can be considered as being the impact of different concepts on the users' preferences. Rather than these weights being provided by the user, in this paper, we attempt to calculate them by observing the users' behaviour.

Ghani and Fano [2002] present their case study of a recommender system based on a custom-built knowledge base of product semantics. The focus within the paper is on generating "soft" attributes from online marketing text, describing the products browsed, and using them to generate cross category recommendations. The profiles used are defined using these derived attributes as opposed to product attributes as is the case in this paper.

6. Conclusions and Future Work

This paper presents only the first steps towards the use of a domain ontology for discovering useful insights into user

behaviour on the web. Our approach aims to calculate the impact of different concepts defined within the ontology on the users navigational behaviour (selection of items). In our experiments we found that the behaviour of different visitors is affected to varying degrees by the various concepts within the ontology. We suggested that these impact values can be used to more accurately determine distance between different users as well as between user preferences and other items on the web site, two basic operations carried out in content and collaborative filtering based recommendations.

The research presented in this paper can be extended in a number of ways. Firstly, we need to evaluate the benefit of using the impact values within the recommendation process. In this paper we only provide an intuitive justification for doing so and presented some evidence to support our claim using a real-world data set.

Secondly, we need to investigate the calculation of similarity further. Haase *et al.* define four layers of an ontology at which similarity of individual concept instances can be calculated [2004]. In the future we intend to investigate similarity at the graph layer. For example, actors and directors are both structurally related to the movie concept which in turn is related to genre. Hence, both actors and directors may be compared using this relation based on the genre profile of the movies they have acted/directed.

Thirdly, in the experiments reported on in this paper, we did not take the time spent by the user on a page. The calculation of impact of a concept lends itself in quite a straightforward way to the incorporation of time spent on a page.

Fourthly, textual information such as the movie plot as well as reviews can be used to identify new keywords which could be used to refine the genre hierarchy which has currently only got a maximum depth of three.

Finally, the role of context needs to be investigated further. While different behaviours in different visits by the same visitor suggest the existence of the role of context within user behaviour, there is little research that has been carried out to date within recommendation technologies that attempt to identify and leverage the user context.

References

[Avancha *et al.*, 2001] S. Avancha, S. Kallurkar, and T. Kamdar. Design of Ontology for The Internet Movie Database (IMDb). Semester Project, CMSC 771, 2001.

[Berendt and Teltzrow, 2005] B. Berendt and M. Teltzrow. Addressing Users' Privacy concerns for Improving Personalization Quality: Towards an Integration of User Studies and Algorithm Evaluation, To appear in *Intelligent Techniques in Web Personalisation* (ed.) B. Mobasher and S. S. Anand, 2005.

[Burke, 2002] R. Burke. "Hybrid recommender systems: survey and experiments", *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331-70, 2002.

[Cho and Kim, 2004] Y.H. Cho and J.K. Kim. Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce. *Expert Systems with Applications*, 26(2), pp. 233-246, 2004.

[Dai and Mobasher, 2003] H. Dai and B. Mobasher. A road map to more effective Web personalization: integrating domain knowledge with Web usage mining. In *Proc. of Intl. Conference on Internet Computing - IC'03*, pp58-64, Las Vegas, Nevada, 2003. CSREA Press.

[Ganesan *et al.*, 2003] P. Ganesan, H. Garcia-Molina, and J. Widom. Exploiting hierarchical domain structure to compute similarity. *ACM Transactions on Information Systems*, 21(1), pp. 64-93, 2003.

[Ghani and Fano, 2002] R. Ghani, and A. Fano. Building Recommender Systems using a Knowledge Base of Product Semantics. Accenture Technology Labs, 2002.

[Guarino and Giaretta, 1995] N. Guarino and P. Giaretta. Ontologies and Knowledge Bases: towards a terminological clarification, Towards Very Large Knowledge Bases: *Knowledge Building and Knowledge Sharing*, pp. 25-32, 1995.

[Haase *et al.*, 2004] P. Haase, M. Ehrig, A. Hotho, and B. Schnizler. Personalized Information Access in a Bibliographic Peer-to-Peer System. In *Proc. of Workshop on Semantic Web Personalization*, 1-12, San Jose, CA, 2004. AAAI.

[Hartigan and Wong., 1979] J. Hartigan and M. Wong. Algorithm AS136: A k-means clustering algorithm. *Applied Statistics*, 28, 100-108. 1979.

[Lang., 1995] K. Lang. NewsWeeder: learning to filter net-news. In *Proc. of Intl Conference on Machine Learning*, pp331-9, Tahoe City, CA, 1995. Morgan Kaufmann.

[Middleton *et al.*, 2004] S.E. Middleton, N.R. Shadbolt, and D.C. De Roure. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems*, 22(1), pp. 54-88, 2004.

[Mobasher *et al.*, 2000] B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on Web usage mining. *Comm. of the ACM*, 43(8), pp. 142-151, 2000.

[Mobasher *et al.*, 2002] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. *Data Mining and Knowledge Discovery*, Vol. 6, No. 1, pp. 61-82, 2002.

[Mobasher *et al.*, 2004] B. Mobasher, X. Jin, and Y. Zhou. Semantically Enhanced Collaborative Filtering. In *Proc. of the European Web Mining Forum*, B Berendt et al. (ed.), LNAI (Volume 3209), Springer, 2004.

[Montaner *et al.*, 2003] M. Montaner, B. Lopez, and J.L. de La Rosa. A taxonomy of recommender agents on the Internet. *Artificial Intelligence Review*, vol. 19, no. 4, pp. 285-330, 2003.

[Nichols, 1998] D.M. Nichols. Implicit Rating and Filtering. In *Proc. of the fifth DELOS Workshop on Filtering and Collaborative Filtering*, pp. 31—36, Budapest, Hungary, 1998. ERCIM.

[Niu *et al.*, 2002] L. Niu, X. Yan., C. Zhang, and S. Zhang. Product hierarchy-based customer profiles for electronic commerce recommendation. In *Proc. of 1st Intl. Conference on Machine Learning and Cybernetics*, pp1075-1080, Beijing, China, 2002. IEEE.